Sequence and hydropathy profile analysis of two classes of secondary transporters

Lolkema, JS; Slotboom, DJ

Published in:
Molecular Membrane Biology

DOI:
10.1080/09687860500063324

Molecular Membrane Biology, May–June 2005; 22(3): 177–189

Taylor & Francis
Taylor & Francis Group

Sequence and hydropathy profile analysis of two classes of secondary transporters

JUKE S. LOLKEMA[1] & DIRK-JAN SLOTBOOM[2]

[1]*Molecular Microbiology, Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, and* [2]*Membrane Enzymology, Biomolecular Sciences and Biotechnology Institute, Department of Biochemistry, University of Groningen, Groningen, The Netherlands*

(Received 26 October 2004; and in revised form 21 December 2004)

Abstract
A structural class in the MemGen classification of membrane proteins is a set of evolutionary related proteins sharing a similar global fold. A structural class contains both closely related pairs of proteins for which homology is clear from sequence comparison and very distantly related pairs, for which it is not possible to establish homology based on sequence similarity alone. In the latter case the evolutionary link is based on hydropathy profile analysis. Here, we use these evolutionary related sets of proteins to analyze the relationship between E-values in BLAST searches, sequence similarities in multiple sequence alignments and structural similarities in hydropathy profile analyses. Two structural classes of secondary transporters termed ST[3], which includes the Ion Transporter (IT) superfamily and ST[4], which includes the DAACS family (TC# 2.A.23) were extracted from the NCBI protein database. ST[3] contains 2051 unique sequences distributed over 32 families and 59 subfamilies. ST[4] is a smaller class containing 399 unique sequences distributed over 2 families and 7 subfamilies. One subfamily in ST[4] contains a new class of binding protein dependent secondary transporters. Comparison of the averaged hydropathy profiles of the subfamilies in ST[3] and ST[4] revealed that the two classes represent different folds. Divergence of the sequences in ST[4] is much smaller than observed in ST[3], suggesting different constraints on the proteins during evolution. Analysis of the correlation between the evolutionary relationship of pairs of proteins in a class and the BLAST E-value revealed that: (i) the BLAST algorithm is unable to pick up the majority of the links between proteins in structural class ST[3], (ii) 'low complexity filtering' and 'composition based statistics' improve the specificity, but strongly reduce the sensitivity of BLAST searches for distantly related proteins, indicating that these filters are too stringent for the proteins analyzed, and (iii) the E-value cut-off, which may be used to evaluate evolutionary significance of a hit in a BLAST search is very different for the two structural classes of membrane proteins.

Keywords: *Secondary transporters, hydropathy profile analysis, structural classes, distant relationships, BLAST searches, low complexity filtering*

Introduction

Secondary transporters are integral membrane proteins that are widely spread in nature. All living organisms use secondary transporters to translocate solutes across cell membranes driven by ion gradients. Analysis of a number of bacterial genomes has indicated that up to 7% of the genes may code for secondary transporters [1]. Their wide distribution and high abundance correlates with great amino acid sequence diversity and the transport protein classification (TC) system developed by Saier and coworkers lists 84 different families in the 'Porters' subclass of the 'Electrochemical potential-driven transporters' category [2]. The large number of families of secondary transporters is not likely to reflect as many different structures and mechanisms but the sequence differences between some families are expected to be the result of divergent evolution. In fact, the TC system recognizes the existence of several superfamilies that group distantly related families, like the Major Facilitator Superfamily (MFS), the amino acid/polyamine/organocation superfamily (APC), and the ion transporter superfamily (IT). Recent crystal structures of two proteins from different families in the MFS, the lactose transporter LacY [3] and the glycerol-P/Pi exchanger GlpT [4] of *Escherichia coli*, revealed very similar structures and both support the alternating access mechanism for solute translocation. In contrast, the 3D structures of the drug transporter AcrB [5], the mitochondrial ADP/ATP carrier [6] and a glutamate transporter homologue Glt_{Ph} [7] all show different

Correspondence: Juke S. Lolkema, Molecular Microbiology, Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751NN Haren, The Netherlands. E-mail: j.s.lolkema@rug.nl

ISSN 0968-7688 print/ISSN 1464-5203 online
DOI: 10.1080/09687860500063324

structural organizations. Moreover, the presence of pore-loop structures in the latter transporter, features not observed in the LacY and GlpT proteins, suggests that fundamentally different translocation mechanisms exist for secondary transport [8,9]. Cysteine scanning experiments have revealed the existence of pore-loop structures in other secondary transporter families [10–14].

In an earlier paper we have proposed a procedure to classify families of secondary transporters into structural classes to discriminate between different 3D-structures and, possibly, different mechanisms [15]. The procedure uses the discriminative power of family hydropathy profiles to distinguish between different folds. Hydropathy profiles are able to detect more distant evolutionary relationships between families of membrane proteins than amino acid sequences alone. In a recent publication we presented one of these structural classes, a class of secondary transporters, termed ST[3] [16] consisting of 568 unique sequences predominantly from prokaryotic origin. Most of the characterized transporters transport organic and inorganic anionic substrates while a smaller fraction represents Na^+/H^+ antiporters. Structural class ST[3] includes the families of the IT superfamily in the TC classification system, but in addition families that cannot be shown to be homologous based on sequence alone.

In this contribution, we use the set of related proteins in structural class ST[3] for a comprehensive analysis of the relationship between sequence identity and hydropathy profile similarity. The E-values observed in BLAST searches between pairs of proteins and their evolutionary relationship in terms of belonging to the same subfamily, family, or structural class are analyzed and the frequencies of 'true' and 'false' positives are determined. The effect of filters, which are default 'on' in BLAST searches is also demonstrated. In addition, a new structural class of secondary transporters, termed ST[4], is presented, which allows the analysis of sequence and hydropathy profile relationships between different structural classes.

Methods

A structural class in the MemGen database is a hierarchical structure of families and subfamilies. Subfamilies represent proteins that produce significant pair wise sequence identities in a multiple sequence alignment while families group subfamilies of which the members produce a significant local alignment in a BLAST search. All proteins in a structural class are expected to be evolutionary related and, therefore, have the same global structure. A structural class contains a subset of the entries in the NCBI protein database [http://www.ncbi.nlm.nih.gov/entrez/] that are stored locally in the MemGen database [http://molmic35.biol.rug.nl/memgen/main.htm]. Building of a structural class involves a combination of BLAST searches [17], multiple sequences alignments [18] and hydropathy profile alignments [15] and groups all members of the class present in the NCBI protein database. The selection procedure consists of four steps; the first two steps are iterative procedures using a 'serial' and a 'parallel' BLAST search approach to define the subfamilies in the class, the third step groups the subfamilies in families and the fourth step confirms the assignment of the subfamilies to the class by family hydropathy profile alignment. Details about the procedure and the algorithms that were used have been reported before [15,16] and can be found at our website [http://molmic35.biol.rug.nl/memgen/main.htm].

BLAST searches were run locally to 'freeze' the contents of the NCBI database. At the times indicated in Table I, the nr database was downloaded from the NCBI ftp site [ftp://ftp.ncbi.nih.gov/BLAST/] and formatted using the formatdb executable. Queries were run against the local database using the blastpgp executable. Both the latter programs are made available by the NCBI at the same ftp site. BLAST searches were done with low complexity filtering (LCF) and composition based statistics (CBS) 'off' (unfiltered) or 'on' (filtered). The number of hits to be reported was set to 5000 which retrieves all hits with an Expect value (E-value) equal to or less than 10. The latter is essential to estimate the sensitivity of the search. E-values are not reported as exact values but in a range indicated by the pE value as follows: a pE of 9 represent all values between 1e-009 and 9e-009, and a pE of 40 all values between 1e-049 and 9e-040, etc. Multiple sequence alignments were done using the the command line version of ClustalW for the Windows XP platform that was downloaded from ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw/. The default settings were used.

A slightly modified method was used to select members to be included in the multiple sequence alignments from which the averaged hydropathy profiles and structural divergence score (SDS) values were calculated. First a frequency distribution of the sequence lengths was made using a resolution of 25 residues. Sequences were included that fall in a window of 75 residues around the highest frequency in the distribution. Optimal alignments of averaged hydropathy profiles were computed using gap costs of 0.7 and 0.3 for opening a gap and extending a gap, respectively [15].

Table I. Statistics of structural classes ST[3] and ST[4].

	ST[3]			ST[4]	
	2002[a]	2003[a]	2004[a]	2003[b]	2004[a]
Families	29	32	32	2	2
Subfamilies	48	55	59	7	7
Unique sequences	568	1183	2051	283	399
Typical sequences	376	666	914	115	154
Sources, distribution					
Archaea		58	61	8	9
Bacteria		1047	1334	211	250
Eukarya		74	92	62	74
Other		4	563	0	64
Sources, represented					
Archaea		19	21	7	8
Bacteria		142	178	85	107
Eukarya		24	30	17	22

[a]April; [b]June.

Results

Structural classes ST[3] and ST[4]

Two structural classes were used in the analysis described here. Structural class ST[3], a class of anion transporters and Na^+/H^+ antiporters, which was first defined in [16] and of which an update is given here, and a new class containing bacterial and neuronal glutamate/aspartate transporters termed ST[4]. ST[3] includes the Ion Transporter (IT) superfamily in the transport protein classification (TC) system ([19] and ST[4] includes the DAACS family (TC 2.A.23) [http://tcdb.ucsd.edu/tcdb/]. Structural classes ST[3] and ST[4] may be browsed at http://molmic35.biol.rug.nl/memgen/main.htm.

In a time span of 2 years, the number of unique sequences in structural class ST[3] has increased 4-fold (Table I). Unique sequences in the MemGen database are tagged as 'typical' or 'similar'. Typical sequences have less than 60% sequence identity with any other typical sequence in the database. Similar sequences share more than 60% identity with a typical sequence. The fraction of typicals in structural class ST[3] decreased from 0.66 in 2002, via 0.56 in 2003 to 0.45 in 2004, indicating that the number of closely related sequences increases at a higher pace than the more distantly related sequences.

Structural class ST[4] is much smaller than ST[3] containing 399 unique sequences (Table I). ST[4] contains two families and 7 subfamilies, but 396 of the 399 unique sequences are in one single family, [st401]DAACS, while the other family, [st402]GTL (Glutamate Transporter Like), consists of only 3 sequences. Apparently, the proteins in ST[4] have diverged much less during evolution as compared to the proteins in class ST[3], which is also reflected in the lower fraction of typical sequences (38% in ST[4] *vs.* 45% in ST[3]). In contrast to ST[3], which contains only 4% sequences of eukaryotic origin, ST[4] has 19% eukaryotic proteins. Archaeal proteins are only poorly represented in both classes. Subfamily [st401]DAACS1 contains transporters from all three kingdoms of life (Table II). The

Table II. Overview of the subfamilies in structural class ST[4].

Subfamily	Unique sequences	Kingdom	Selected members	Known substrates
[st401]DAACS1	317	B,A,E	EAAT1	Neural glutamate/aspartate transporters
			AAAT	Eukaryotic neutral aminoacid transporters
			GLTP	Bacterial glutamate transporters
			DCTA	Bacterial C4 dicarboxylate transporters
[st401]DAACS2	30	B		
[st401]DAACS3	19	B,E		
[st401]DAACS4	28	B	SSTT	Bacterial serine transporters
[st401]DAACS5	2	B,A		Transporter/binding protein hybrids
[st402]GTL1	1	B		
[st402]GTL2	2	B		

well-studied bacterial and neuronal glutamate/aspartate transporters are found in this subfamily as well as the eukaryotic neutral amino acid transporters and bacterial dicarboxylate transporters (for a review see [20]). The glutamate transporter homologue from *Pyrococcus horikoshi* of which a crystal structure was recently solved [7] is a member of this subfamily. [st401]DAACS4 contains bacterial serine transporters [21]. Many of the characterized transporters use the Na^+ gradient across the membrane to drive substrate uptake. The remaining subfamilies represent transporters from the bacterial kingdom with the exception of a sequence from the mosquito *Anopheles gambiae* in [st401]DAACS3 and an archaeal protein in [st401]DAACS6. The three sequences in the second family in ST[4], [st402]GTL1 are from organisms from the phylum Spirochaetales, *Borrelia burgdorferi*, *Treponema pallidum*, and *Treponema denticola*.

A new class of hybrid proteins in the [st401]DAACS family

The two proteins in subfamily [st401]DAACS6 in class ST[4], NP615455mace from the archaeon *Methanosarcina acetivorans* and PMIT0791prma from the cyanobacterium *Pyrochlorus marinus*, are unusually long compared to the other prokaryotic members in the class. The hydropathy profiles of the sequences show that each protein consists of an integral membrane protein part with a soluble protein domain fused at its C-terminus (Figure 1A for PMIT0791prma). Alignment of the hydropathy profiles of the membrane domains of the two [st401]DAACS6 proteins and the [st401]DAACS1 subfamily profile revealed high similarity in the N-terminal halves but significant deviation in the C-terminal part (see Figure 1B for NP615455mace), suggesting that only the N-terminal parts share a similar fold. This would be in line with the local alignments obtained in the BLAST searches that all cover at least the N-terminal parts of the sequences. An extra hydrophobic segment is present just before the soluble domains of the two [st401]DAACS6 proteins. The soluble domains are homologous to extracellular substrate binding proteins of ABC transport systems. Submission of the soluble domains including the preceding hydrophobic segment to the signal peptide prediction server SignalP [http://www.cbs.dtu.dk/services/SignalP; 22,23] results in the prediction of cleavable signal sequences with high confidence (Gram-positive Hidden Markov model) suggesting that the genes code for a secondary transporter that functions with an extra cellular binding protein.

ST[3] and ST[4] have different folds

In the MemGen classification, the criterion for structural similarity of two (sub)families is based on a statistical analysis of the averaged hydropathy profiles of both families and the optimal alignment of the two averaged profiles [15]. The method relies heavily on the quality of the multiple sequence alignment that is used to calculate the averaged hydropathy profile and the divergence of the individual profiles. Table III summarizes some of the properties of the multiple sequence alignments of the typical sequences of the subfamilies in ST[3] and ST[4]. The S-test compares the difference between two averaged subfamily profiles after optimal alignment (PDS; Profile Difference Score) with the divergence of the individual profiles within each of the two averaged profiles (SDS). A structural relationship between subfamilies is evident when the value for the S-test is around 1 or lower. The test requires that the subfamilies to be compared contain enough sequences to compute a statistically meaningful averaged hydropathy profile and SDS value. Currently, the minimal number of sequences in a subfamily is set to 8 and this criterion is fulfilled by 32 and three subfamilies in ST[3] and ST[4], respectively (Table III). The 32 subfamilies of ST[3] are all grouped into the same structural class by the S-test (not shown; see [14] for the previous version of ST[3] and Figure 2A for an example). Profile alignment of the three subfamilies of ST[4] resulted in S-test values of 0.62 ([st401]DAACS1 and [st401]DAACS2), 0.62 ([st401]DAACS1 and [st401]DAACS3), and 1.16 ([st401]DAACS2 and [st401]DAACS3), indicating similar hydropathy profiles and confirming their assignment to one and the same structural class.

Alignment of the averaged hydropathy profiles of the 3 subfamilies in ST[4] with the 32 subfamilies of ST[3] listed in Table III revealed significant differences between the profiles. The averaged values for the S-tests for the [st401]DAACS1, [st401]DAACS2, and [st401]DAACS3 subfamilies and the 32 ST[3] subfamilies were 2.3 ± 0.5, 2.8 ± 0.8, and 1.9 ± 0.5, respectively. Figure 2B shows the optimal alignment for the [st401]DAACS1 and [st310]ATO1 families as an example. Using our criterion for structural similarity, this clearly indicates that ST[3] and ST[4] represent classes of membranes proteins with different folds. Four out of the 96 hydropathy profile alignments between the 3 ST[4] and 32 ST[3] families resulted in S-test values lower than 1.2 suggesting structural similarity and should be considered false positives. Such alignments are artifacts of the alignment algorithm when the profiles are of considerably different length and

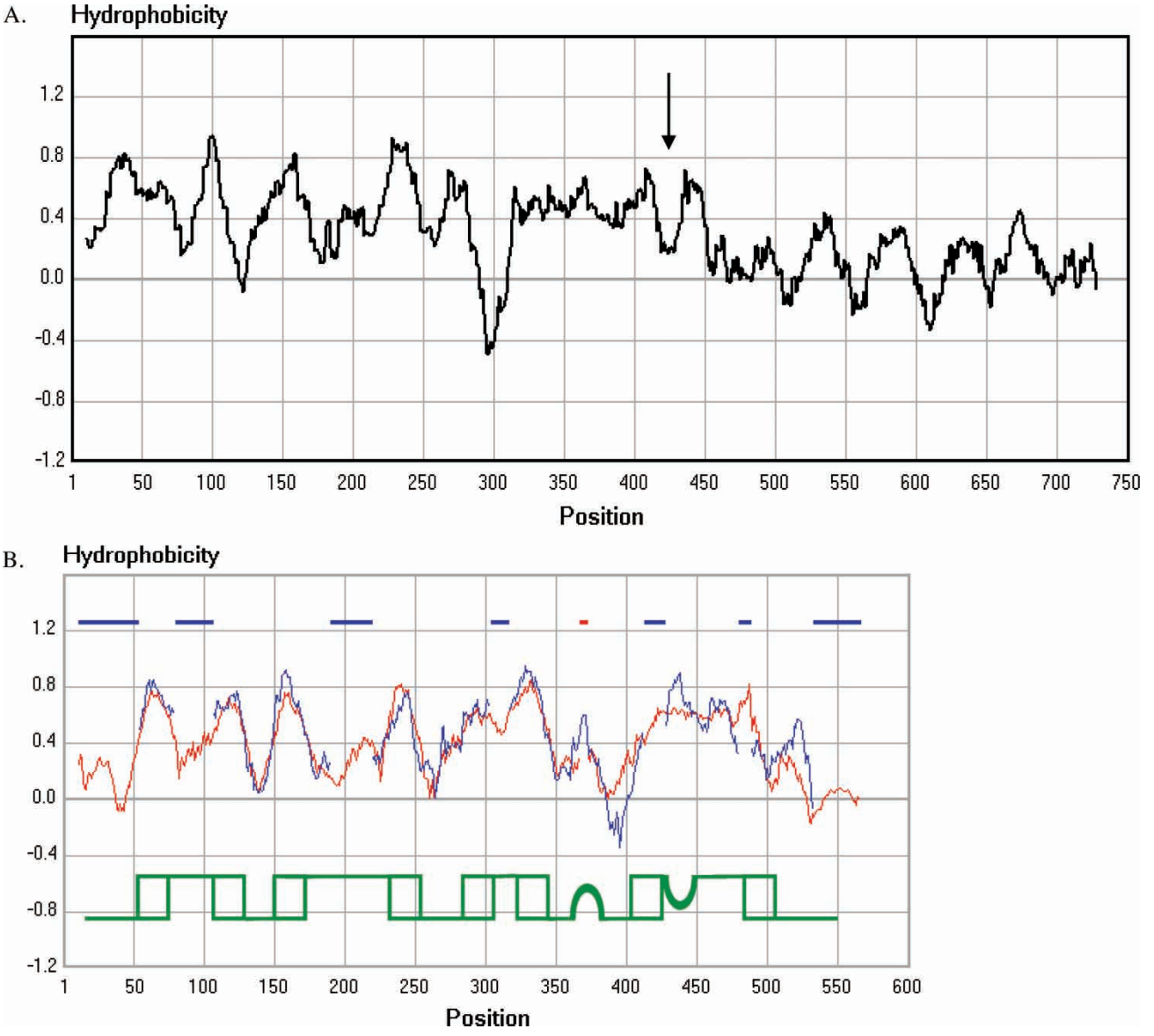

Figure 1. The [st401]DAACS6 subfamily. (A) Hydropathy profile of PMIT0791prma. The arrow indicates the end of the transporter domain and the beginning of the substrate binding domain. (B) Optimal alignment of the hydropathy profile of the transporter domain of NP615455mace (black) and the averaged profile of [st401]DAACS1 (gray). The PDS of the alignment was 0.136. Bars at the top indicate positions where gaps were introduced in the corresponding profiles. Below, the membrane topology model of the [st401]DAACS1 transporters based on the 3D structure of the *Pyrococcus horikoshii* glutamate transporter homologue is indicated. Boxes represent transmembrane segments, while arcs in loop regions represent reentrant loops. The bottom of the scheme corresponds to the cytoplasm. This figure is reproduced in colour in *Molecular Membrane Biology* online.

containing different numbers of hydrophobic regions. False positives are immediately evident upon visual inspection of the alignments.

Correlation between E-value and evolutionary relationship

The 914 and 154 typical sequences in structural classes ST[3] and ST[4], respectively, provide sets of proteins that are evolutionary related. BLAST searches against a database of all typicals in a class and using all typicals as queries would return a maximum of 834,482 hits for ST[3] and 23,562 for ST[4] if all typicals would 'see' one another. Using a cut-off for the E-value of 10, the BLAST searches of the NCBI database detected 17.4% of the possible hits between the typicals in ST[3] and 98.9% between the typicals in ST[4]. The high percentage for ST[4] is a manifestation of the low divergence of the sequences in ST[4] as noted earlier. In the MemGen classification system evolutionary links between proteins have different scopes. The links

Table III. Properties of subfamilies in ST[3] and ST[4].

Subfamily	Typicals[a]	Residues[b] range	Pair wise SI (%)[c] range	Pair wise SI (%)[c] median	SDS[d]
[st301]MeCit1	14/14		23–61	35	0.107
[st302]ArsB1	14/18		17–58	26	0.124
[st302]ArsB2	15/25	421–468	18–56	29	0.120
[st302]ArsB3	13/13		20–48	26	0.132
[st302]ArsB4	11/15	395–421	21–55	29	0.129
[st303]AIT1	21/29	467–517	18–59	33	0.116
[st303]AIT2	29/64	432–520	18–59	26	0.134
[st303]AIT3	32/52	574–621	20–60	26	0.145
[st304]GNT1	32/42	437–461	21–61	31	0.120
[st304]GNT2	12/12		38–61	46	0.099
[st305]DCUA1	13/14		25–57	39	0.100
[st307]DCTM1	116/173	402–463	18–60	28	0.126
[st307]DCTM2	15/39	652–712	23–52	33	0.135
[st309]DCUC1	9/9		21–47	31	0.115
[st310]ATO1	16/16		24–60	33	0.114
[st311]AITB1	12/12		27–56	39	0.124
[st312]NHAC1	37/43	446–491	20–58	31	0.119
[st312]NHAC2	19/27	460–577	22–61	30	0.120
[st312]NHAC3	18/21		26–59	39	0.122
[st313]AITC1	18/23	465–518	20–61	30	0.124
[st314]AITD1	25/25		27–60	38	0.111
[st315]AITE1	17/30	541–597	19–60	25	0.128
[st316]NHAD1	9/10		24–59	29	0.127
[st317]AITF1	8/8		18–33	24	0.132
[st318]AITG1	10/10		20–59	27	0.124
[st320]AITI1	9/9		25–60	32	0.121
[st321]AITJ	8/8		35–57	41	0.114
[st324]GLTS1	19/19		26–57	35	0.109
[st325]2HMCT1	22/23		23–60	33	0.122
[st325]2HMCT2	16/16		25–60	36	0.120
[st326]2HCT1	15/16		22–58	30	0.120
[st330]AITP2	8/8		18–58	29	0.102
[st401]DAACS1	67/119	395–458	19–60	28	0.128
[st401]DAACS2	14/14		28–58	42	0.103
[st401]DAACS3	10/11		21–52	39	0.102

[a]The number of typicals in the subfamily included in the analysis. [b]Size window used to select sequences. [c]Indicated were the lowest and highest sequence identity (SI) between pairs in the subfamily and the median of the pair wise sequence identity distribution. [d]Structural Divergence Score.

may be between proteins of the same subfamily (scope: subfamily), between proteins in different subfamilies of the same family (scope: family) or between proteins in different families of the same class (scope: class). The evolutionary distance increases in the order subfamily < family < class. The classification allows us to analyze the relationship between the evolutionary distance and the E-value obtained in the BLAST searches. Figure 3A and 4A give the cumulative frequency distribution of the E-values of the hits between the typicals in ST[3] and ST[4], respectively, categorized according to scope. In ST[3], the number of observed hits between proteins of scope subfamily is 100%, of scope family 86.5%, and of scope class only 8%. Clearly, the BLAST algorithm does not see the majority of the links with scope class in ST[3]. In contrast, in ST[4] the fraction of the links with scope class observed at an E-value equal to or smaller than 10 amounts to 62% (Figure 4A), which is another manifestation of the much lower divergence of the sequences in ST[4] compared to ST[3].

The effect of filtering in BLAST searches

In the analysis above, the BLAST searches were not filtered for low complexity regions (LCF) and composition based statistics (CBS) was not applied. The effect of these filters on the frequency distribution of the E-values of the hits between the typicals in ST[3] and ST[4] is shown in Figure 3B and 4B, respectively. As expected, filtering had little effect on the hits with scope subfamily. The number of observed hits was hardly affected and the most

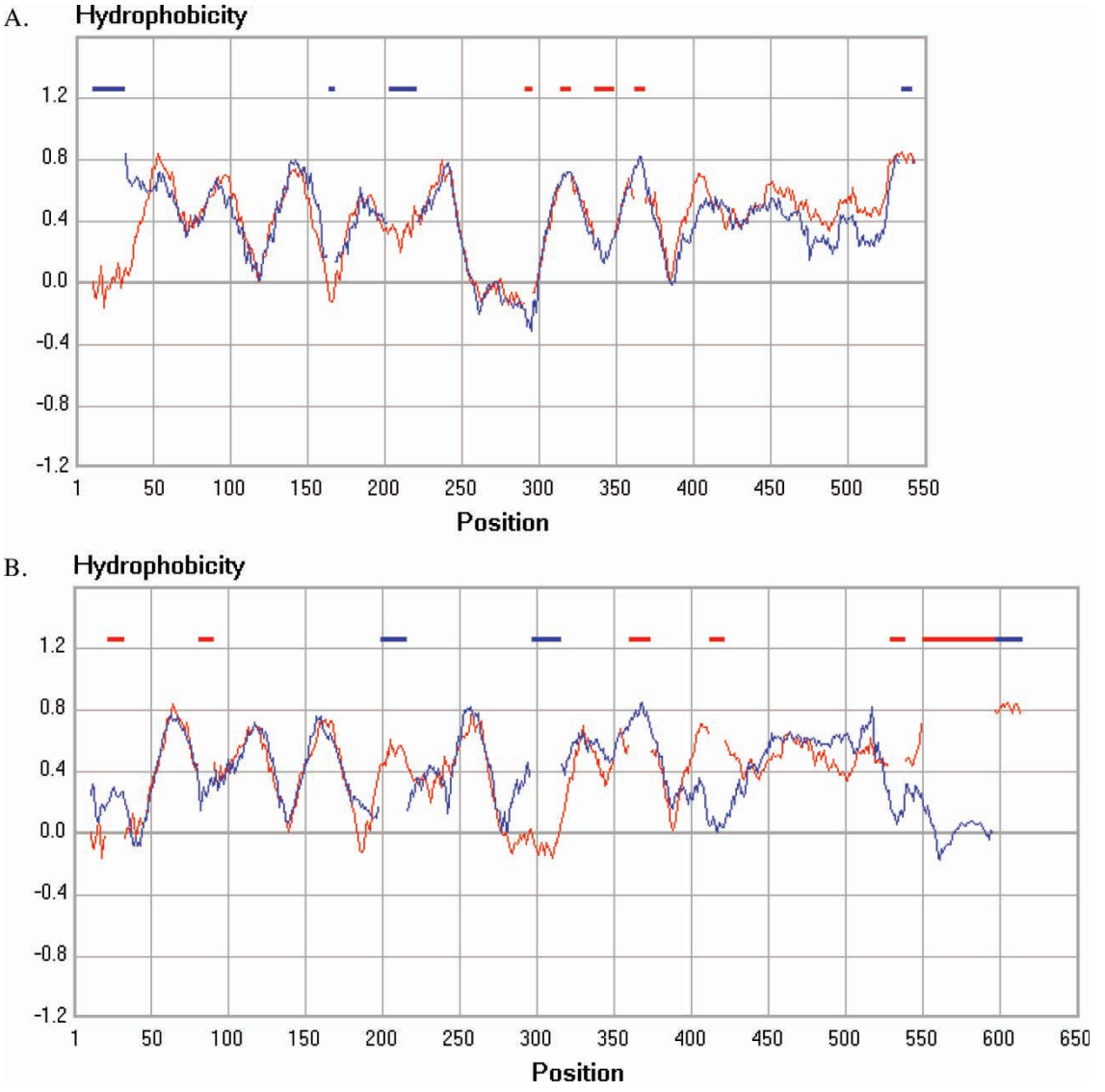

Figure 2. Optimal alignment of the family hydropathy profiles of [st310]ATO1 in ST[3] and [st311]AITB1 in ST[3] (A) and [st401]DAACS1 in ST[4] (B). The S-test for the alignments were 1.13 (A) and 1.73 (B), respectively. In filtered BLAST searches, no hits were observed between the members of the [st310]ATO1 family and the members of the [st311]AITB1 and [st401]DAACS1 families. The [st310]ATO1 profile is given in black, the [st311]AITB1 and [st401]DAACS1 profiles in gray. Bars in the top of the figure indicate the positions of gaps introduced by the alignment procedure in any of the two profiles. This figure is reproduced in colour in *Molecular Membrane Biology* online.

prominent difference was a general shift to higher E-values. The effect of filtering was more significant on hits with scope family, especially in class ST[3]. The percentage of observed hits dropped from 86.5% to 49%; more than half of the evolutionary links with scope family were not observed in the filtered BLAST searches. For ST[4] the effect was smaller, the number of observed hits dropped from 93% to 75%. The most dramatic effect of filtering was observed on hits between proteins with scope class. The number of observed hits in ST[3] dropped by a factor of 10 to 0.8% and in ST[4] by a factor of 6 to 11%.

True and false positive BLAST hits

The 914 typical sequences in ST[3] hit a total of 1,023,657 entries in unfiltered BLAST searches of the NCBI protein database and the 154 typical proteins in ST[4] a total of 139,113. The hits include proteins of the same class (true positives) and hits with proteins that are not a member of the class to which the query belongs (false positives). Figure 5A and 5C give the (absolute) frequency distributions of the E values of the hits with scope class and the hits with proteins not in the same class in ST[3] and ST[4], respectively. The distributions

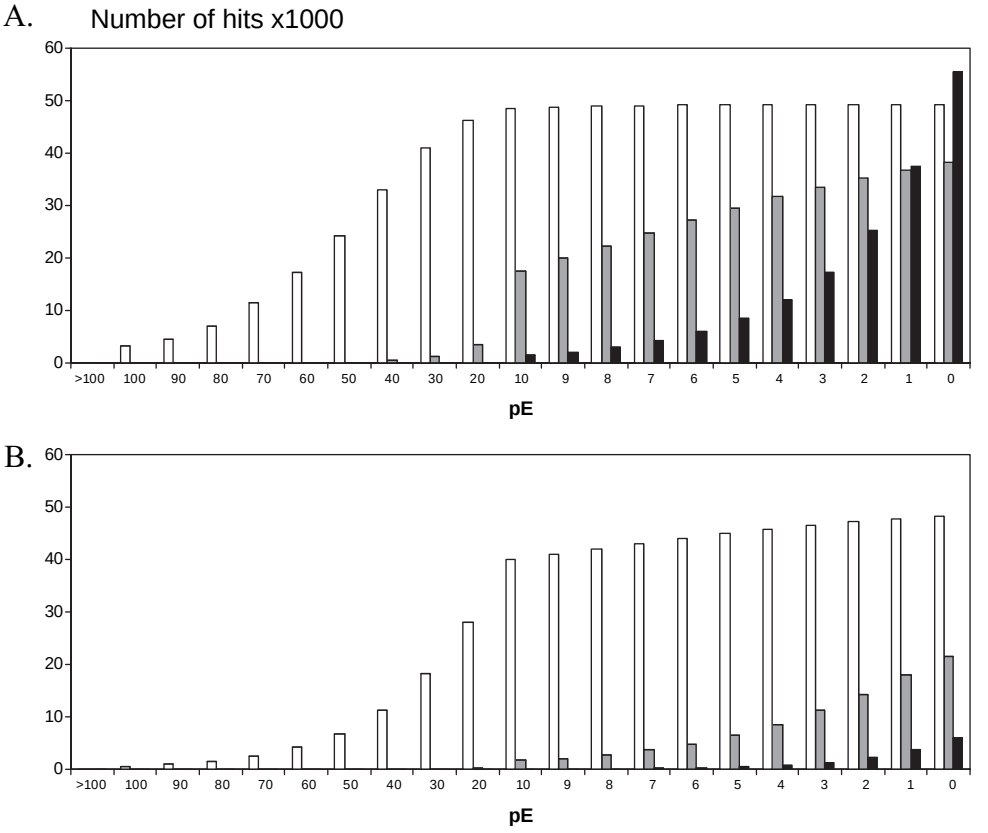

Figure 3. Cumulative distribution of hits between typical sequences in class ST[3] over E-values following BLAST searches using all typical sequences in ST[3] as query. The distribution of the hits with other typicals in ST[3] of scope 'subfamily' (open bars), scope 'family' (grayed bars), and scope 'class' (filled bars) is shown. BLAST searches were performed with the low complexity filter and composition based statistics not set (A; unfiltered) or set (B: filtered). E-values were grouped in the ranges indicated by pE (see Methods).

give a measure of the significance of a hit with a distantly related protein. In ST[3], the first false positives show up in the E-value range between 1e-5 and 9e-5 (pE = 5, see legend), but these are significantly out numbered by the hits inside the class. At a pE of 1 the chance of hitting a false positive is about 50% and at a pE of 0 about 75%. Filtering of the BLAST searches significantly improves the specificity of the search (Figure 5B). The chance of hitting a false positive drops to 10% at a pE of 1 and to below 30% for hits with a pE of 0. The price that is paid for the improved specificity is that the number of observed class members with scope 'class' is decreased 10-fold (see above).

The same analysis for ST[4] results in a dramatically different view (Figure 5C,D). The chance of a hit being a false positive at the lower pE values is dramatically higher for a query in ST[4] than in ST[3]. The chance of hitting a false positive at a pE of 1 is 96%, and at a pE of 0 the chance increases even up to 99%. Again, filtering of the BLAST searches gives significant improvement but the chances of a hit being a true positive are still only in the order of 20% at pE values of 1 and 0.

The fraction of false positive hits of the typical proteins in class ST[3] that corresponds to proteins in ST[4] was slightly less than 1% in an unfiltered search (1627 out of a total of 170358 false positives). There was no correlation between the false positives between the two classes in the BLAST searches and those in the family hydropathy profile alignments (see above), indicating that the mechanism for the appearance of the false positives is not related. The relative distribution over pE values of the hits to ST[4] proteins and those to all false positives was more or less the same, indicating that there is no specific preference for or against hitting a ST[4] protein (not shown). Filtering of the BLAST searches reduced the number of hits with ST[4] proteins 200-fold while the number of hits with all proteins outside ST[3] was reduced by a factor of 60. Therefore, filtering selects against hits to proteins in ST[4]. This once more supports the notion that ST[3] and ST[4] truly represent classes of evolutionary unrelated proteins.

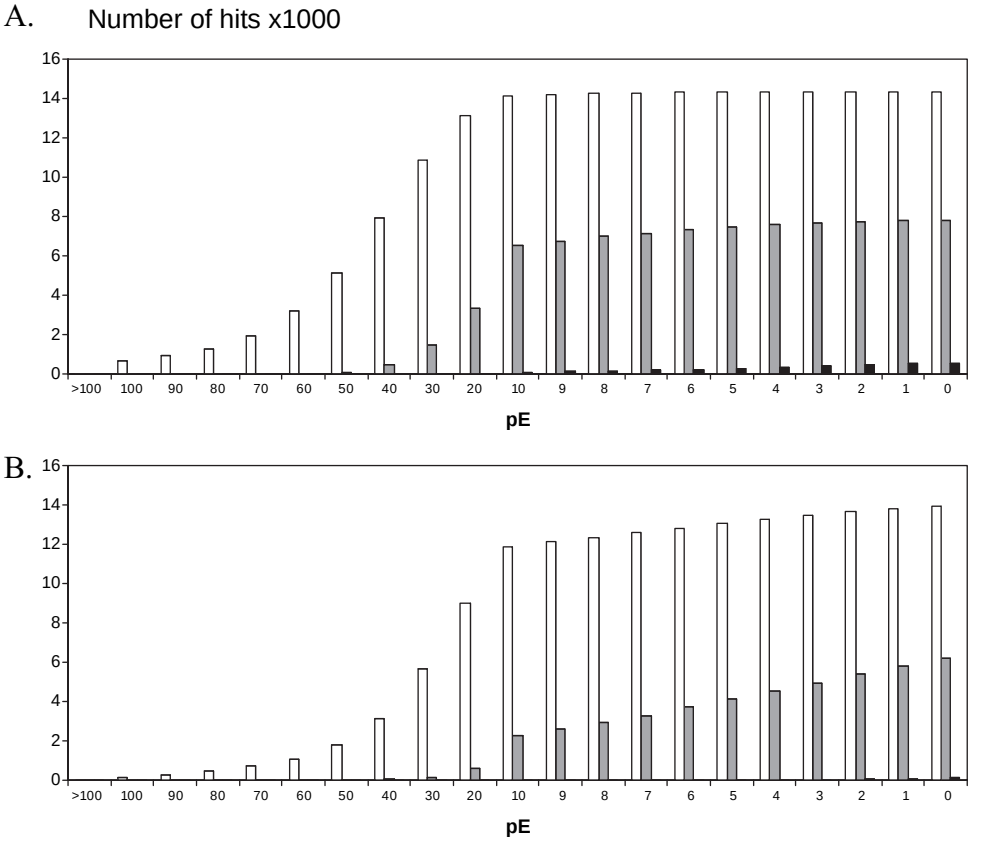

Figure 4. Cumulative distribution of hits between typical sequences in class ST[4] over E-values following BLAST searches using all typical sequences in ST[4] as query. The distribution of the hits with other typicals in ST[4] of scope 'subfamily' (open bars), scope 'family' (grayed bars), and scope 'class' (filled bars) is shown. BLAST searches were performed with the low complexity filter and composition based statistics not set (A; unfiltered) or set (B: filtered). E-values were grouped in the ranges indicated by pE (see Methods).

Discussion

The Sargasso Sea sequencing project

The distribution of the unique sequences over biological sources indicates that structural class ST[3] is a typical bacterial class of secondary transporters [16]. In the last year, a major contribution was made to the NCBI protein database by a genome shotgun sequencing project of environmental samples collected at sites in the Sargasso Sea near Bermuda [24]. The sampling technique selected for microbial populations and in the latest update, sequences from this project constitute up to 26% of the sequences in ST[3], i.e., 26% of the sequences are from unknown microbial origin. An unusual high fraction of these sequences are found in one particular family, [st307]DCTM (TC# 2.A.56), which is one of the major families in ST[3]. Of all sequences in ST[3], 27% are found in the [st307]DCTM family, while 52% of the 'Sargasso Sea' sequences are found in this family. The [st307]DCTM family contains the so-called TRAP (tripartite ATP-independent periplasmic) transporters, which are not conventional secondary transporters but part of a larger complex consisting of two membrane proteins that are fused in some members [25,26]. Similar to ABC transporters, the transport systems require a periplasmic binding protein but unlike the ABC transporters electrochemical ion gradients and not ATP hydrolysis is the driving force for the transport reaction. The involvement of extra cellular binding proteins may make them especially suitable for marine environments where nutrient concentrations are likely to be very low. Other well-populated families in ST[3] are only poorly represented in the Sargasso Sea, like [st301]MeCit, [st302]ArsB, and [st326]2HCT. Transporters from structural class ST[4] are less abundant in the Sargasso Sea (17% and 26% for ST[4] and ST[3], respectively).

Binding protein dependent secondary transporters

The TRAP transporters in the [st307]DCTM family in structural class ST[3] appear to have their counterparts in the [st401]DAACS6 subfamily in ST[4]. Fused to the C-terminus of NP615455mace from the archaeon *Methanosarcina acetivorans* and

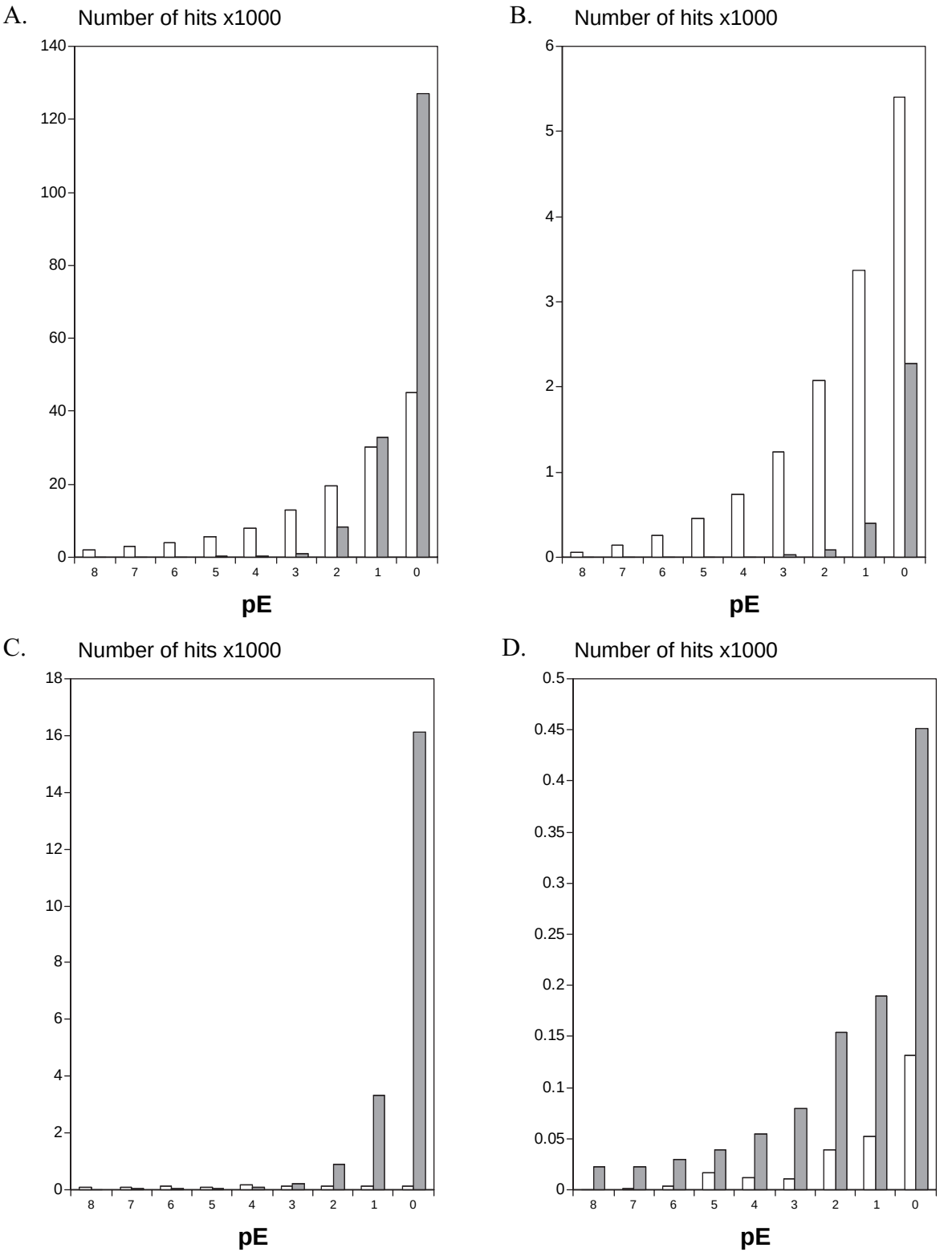

Figure 5. True and false positive BLAST hits. Distribution of the hits of all typicals in ST[3] (A,B) and ST[4] (C,D) in unfiltered (A,C) and filtered (B,D) BLAST searches with proteins in the same class of scope 'class' (open bars) and with proteins not in the same class (closed bars). E-values were grouped in the ranges indicated by pE (see Methods).

PMIT0791prma from the cyanobacterium *Pyrochlorus marinus* is a soluble domain that is homologous to substrate binding proteins commonly found in ABC transport systems. The fusion is similar to a recently described group of ABC transporters that have their substrate binding protein

fused to the C-terminal end of the translocator subunits [27]. The binding protein is fused via its protein anchor ('the signal sequence') to the cytoplasmic C-terminus of the translocator domain, which is consistent with a cytoplasmic localization of the C-terminus of the [st401]DAACS6 transporters. The presence of a cleavage site immediately downstream of the C-terminal transmembrane segment, suggests that the binding protein is dissected from the transporter upon maturation.

In the recently solved crystal structure of the *Pyrococcus horikoshii* glutamate transporter homologue [7], a member of [st401]DAACS1, the six transmembrane segments in the N-terminus form a rim around the substrate binding and translocation site formed by the C-terminal half in a trimeric organization. The translocation sites are formed by two trans membrane segments and two reentrant loops formed by helical hairpins that enter the membrane from opposite sides. It is difficult to see how the additional transmembrane segments formed by the signal sequences of the [st401]DAACS6 proteins could be accommodated in this structure and it is likely that they are located at the periphery of the molecule Another possibility is that the C-teminal halves of the transporter domains of the [st401]DAACS6 proteins are structurally dissimilar to the other DAACS subfamilies. This possibility is supported by hydropathy profile analysis that indicates similar folds in the N-terminal halves but dissimilar fold in the C-terminal halves. Moreover, sequence motifs in the C-terminal halves, among which the serine-rich motif characteristic for one of the reentrant loops [20] are missing in the NP615455mace and PMIT0791prma sequences, a feature that is also observed in the sequences in family [st402]GTL, the second family in ST[4] (not shown).

Family hydropathy profiles

A basic assumption in the MemGen classification scheme is that the global folding of a family of integral membrane protein is represented by a unique family hydropathy profile (the family profile). A parameter is defined for each family, the SDS, which measures the difference between the individual hydropathy profiles of the members and the family profile to allow for discrimination between different folds in a numerical procedure (S-test). In practice, both family profile and the value of the SDS are approximated by constructing an averaged hydropathy profile of a multiple sequence alignment of the available members of a family and, consequently, they will deviate from the genuine family profile and SDS value of the fold. Optimal approximation requires that the multiple sequence alignment contains enough members and sequence diversity. Currently, we use a minimum of 8 members with pair wise sequence identities between 20 and 60% while the median of the pair wise sequence identity distribution provides a measure of the composition of the family. Problems are related to sequence errors, regions with low sequence identity, and variations between members such as extra domains, large loop regions and large insertions/deletions. These result in badly defined family profiles and relatively high SDS values. As more and more sequences appear in the databases, a more restrictive selection for the multiple sequence alignment is possible. In the latest analysis of the MemGen database, a selection was made for the large subfamilies based upon a narrow window of sequence lengths (Table III, column 3), which solves a number of the problems mentioned above. As a result, the SDS value for the families went down, which results in a stricter criterion for structural similarity when comparing family profiles. Nevertheless, within class ST[3] many values for the S-test were found to be lower, which has to mean that the new family profiles were better defined.

Differences between ST[3] and ST[4]

The statistical analysis of family hydropathy profiles of (sub)families in the two structural classes of secondary transporters analyzed here, ST[3] an ST[4] demonstrated that the proteins in the two classes have different folds and are not evolutionary related. It is likely that the two classes have originated from different ancestral proteins, but that they have evolved towards similar functions (i.e., secondary transport). Analysis of the BLAST results of representative (typical) sequences in the two classes shows that a large fraction of the evolutionary links between the proteins in class ST[3] is not detected by the BLAST algorithm, while this is much better for class ST[4]. The difference is related to the much lower divergence of the sequences in the latter class. Either the constraints on the structure and/or function were stricter, or the common ancestor arose much later in time. The latter explanation is less likely since members of ST[4] are found both in bacteria and archaea, suggesting that the fold had already evolved in the common ancestor of the domains of life.

Specificity and sensitivity of BLAST searches

The BLAST algorithm produces a list of proteins (the hits) and (local) alignments with the query sequence in decreasing order of similarity to the

query sequence, corresponding to increasing E-values. By adding structural information contained in hydropathy profiles to the sequence analysis by BLAST our method discriminates between true and false positives in BLAST searches of membrane proteins which improves the detection of distantly related membrane proteins. Similar approaches have been explored by others [28,29]. The structural classes constructed in this way group distantly related sequences and can be used, in retrospect, to analyze the sensitivity and specificity of the BLAST searches and the effect of filtering on these parameters.

The BLAST algorithm sees essentially all the hits between typicals with scope 'subfamily' in a class, i.e., all members of subfamilies which have pair wise sequence identities $>20\%$, even after filtering. Also, the majority of the links with scope 'family' were observed (87% for ST[3] and 94% for ST[3]), but here, the default LCF and CBS settings are beginning to ask their toll on the sensitivity of the BLAST search, especially in ST[3] where the fraction of observed links drops to below 50%. The BLAST algorithm performed very poorly in detecting typical-typical links with scope 'class' in ST[3] (8% observed hits) but substantially better in ST[4] (62%). Filtering has a devastating effect on this category by reducing the sensitivity 10-fold and 6-fold for ST[3] and ST[4], respectively. The filters appear to be too strong for distantly related sequences of these types.

The specificity of the BLAST search gives the chance that a hit with a particular E-value is evolutionary significant. Using the typicals in the two classes as queries, we have compared the number of hits with the most distantly related sequences in the same class (scope 'class'; true positives) with the hits with sequences not in the class (false positives). The most remarkable result is that the significance of a hit at a certain E-value is very different for ST[3] and ST[4] (Figure 5). At E-values between 1e-1 and 9e-1, the ratio of 'true' to 'false' is 1:1 and 1:20 for ST[3] and ST[4], respectively. It follows that a threshold E-value for evolutionary significance cannot be given in general. The mechanism behind this is most likely that in a huge database like the NCBI protein database, each query, more or less, picks up at random the same number of false positives. For ST[3] and ST[4] these numbers were 186 and 135 hits per typical, respectively. Then, the specificity is determined by the number of true positives at a certain E-value. If the divergence of a class is small like is the case for ST[4], there will be relatively few hits with high E-value and the specificity will be low. The specificity depends on the divergence of the sequences in the class.

Filtering strongly improves the specificity of the search. The ratio's of 'true' to 'false' change to 8:1 and 1:4 for ST[3] and ST[4], respectively. Clearly filtering selects against the false positives. The specificity increases and at the same time the sensitivity decreases.

Acknowledgements

DJS was supported by a long-term fellowship from the Human Frontier Science Program.

References

[1] Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr. Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J Mol Biol 2000;301:75–100.
[2] Saier MH Jr. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 2000;64:354–411.
[3] Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S. Structure and mechanism of the lactose permease of *Escherichia coli*. Science 2003;301:610–615.
[4] Huang Y, Lemieux MJ, Song J, Auer M, Wang D-N. Structure and mechanism of the glycerol-3-P transporter of *Escherichia coli*. Science 2003;301:616–620.
[5] Murakami S, Nakashima R, Yamashita E, Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. Nature 2003;419:587–593.
[6] Pebay-Peyroula E, Dahout-Gonzalez C, Kahn R, Trezeguet V, Lauquin GJ, Brandolin G. Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. Nature 2003;426:39–44.
[7] Yernool D, Boudker O, Jin Y, Gouaux E. Structure of a glutamate transporter homologue from *Pyrococcus Horikoshii*. Nature 2004;431:811–818.
[8] Slotboom DJ, Konings WN, Lolkema JS. Glutamate transporters combine transporter- and channel-like features. TIBS 2001;26:534–539.
[9] Kanner B, Borre L. The dual-function glutamate transporters: Structure and molecular characterisation of the substrate-binding sites. Biochim Biophys Acta 2002;1555: 92–95.
[10] Qiu Z, Nicoll DA, Philipson KD. Helix packing of functionally important regions of the cardiac Na(+)-Ca(2+) exchanger. J Biol Chem 2001;276:194–199.
[11] Iwamoto T, Uehara A, Imanaga I, Shigekawa M. The Na^+/Ca^{2+} exchanger NCX1 has oppositely oriented reentrant loop domains that contain conserved aspartic acids whose mutation alters its apparent Ca^{2+} affinity. J Biol Chem 2000;275:38571–38580.
[12] Nicoll DA, Ottolia M, Lu L, Lu Y, Philipson DK. A new topological model of the cardiac sarcolemmal Na^+-Ca^{2+} exchanger. J Biol Chem 1999;274:910–917.
[13] Lambert G, Forster IC, Stange G, Kohler K, Biber J, Murer H. Cysteine mutagenesis reveals novel structure-function features within the predicted third extracellular loop of the type IIa Na(+)/P(i) cotransporter. J Gen Physiol 2001;117: 533–546.
[14] Sobczak I, Lolkema JS. Alternating access and a pore-loop structure in the Na^+-citrate transporter CitS of Klebsiella pneumoniae. J Biol Chem 2004;279:31113–31120.

[15] Lolkema JS, Slotboom DJ. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. Mol Membr Biol 1998;15:33–42.
[16] Lolkema JS, Slotboom DJ. Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis. J Mol Biol 2003;327:901–909.
[17] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
[18] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
[19] Prakash S, Cooper G, Singhi S, Saier MH Jr. The ion transporter superfamily. Biochim. Biophys Acta 2003;1618:79–92.
[20] Slotboom DJ, Konings WN, Lolkema JS. Structural features of the glutamate transporter family. Microbiol Molec Biol Rev 1999;63:293–307.
[21] Ogawa W, Kim YM, Mizushima T, Tsuchiya T. Cloning and expression of the Na^+-coupled serine transporter from Escherichia coli and characteristics of the transporter. J Bacteriol 1998;180:6749–6752.
[22] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
[23] Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6), AAAI Press, Menlo Park, California, 1998, pp. 122–130.
[24] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004;304:58–60.
[25] Rabus R, Jack DL, Kelly DJ, Saier MH Jr. TRAP transporters: an ancient family of extracytoplasmic solute-receptor-dependent secondary active transporters. Microbiology 1999;145:3431–3445.
[26] Kelly DJ, Thomas GH. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. FEMS Microbiol Rev 2001;25:405–424.
[27] van der Heide T, Poolman B. ABC transporters: One, two or four extracytoplasmic substrate-binding sites? EMBO Rep 2002;3:938–943.
[28] Geourjon C, Combet C, Blanchet C, Deleage G. Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci 2001;10:788–797.
[29] Hedman M, Deloof H, von Heijne G, Elofsson A. Improved detection of homologous membrane proteins by inclusion of information from topology predictions. Protein Sci 2002;11:652–658.

This paper was first published online on prEview on 21 April 2005.